Intimate Evolution of Proteins : Proteome Atomic Content Correlates with Genome Base Composition.

Peggy Baudouin-Cornu *, Katja Schuerer [§], Philippe Marlière[†]

and Dominique Thomas *

* *Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique*

91 198 Gif sur Yvette Cedex, France

[§] *Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France*

[†] *Evologic SA, 89 rue Henri Rochefort, 91000 Evry, France*

Corresponding author :
Dominique Thomas
Centre de Génétique Moléculaire, CNRS
91 198 Gif-sur-Yvette, France
tel : 33 1 69 82 32 33
FAX : 33 1 69 82 43 72
e-mail : thomas@cgm.cnrs-gif.fr

Abstract

Discerning the significant relations that exist within and among genome sequences is a major step towards the modelling of biopolymer evolution. Here we report the systematic analysis of the atomic composition of proteins encoded by organisms representative of each kingdoms. Protein atomic contents are shown to vary largely among species, the larger variations being observed for the main architectural component of proteins, the carbon atom. These variations apply to the bulk proteins as well as to subsets of ortholog proteins. A pronounced correlation between proteome carbon content and genome base composition was further evidenced, high G+C genome content being related to low protein carbon content. The generation of random proteomes and the examination of the canonical genetic code provide arguments for the hypothesis that natural selection might have driven genome base composition.

Introduction

Comparative analyses of complete genome sequences were anticipated to reveal the molecular bases of biodiversity as well as to increase our comprehension of the constraints that shaped protein composition and structure during the natural history of living organisms. However, while comparative genomics highlighted genome structure plasticity and strengthened the role of lateral gene transfer during the natural history of living organisms [1, 2], less progress was accomplished in understanding adaptive evolution at the molecular level and how it contributes to changes in proteins.

Studies devoted to protein evolution mostly used comparative analyses of protein primary sequences and attempted to identify which rules govern conservation, substitutions and deletions of amino acids that are observed between proteins [3, 4, 5]. By focusing on amino acid composition, such studies did not address the possibility that the evolution of proteins, and more generally of biopolymers, might have been shaped at a more intimate level : the atomic level. It is worth to note however that, among the constitutive elements of proteins, carbon, nitrogen and sulphur atoms are themselves subject to geochemical cycles at the surface of the Earth [6]. As a consequence, large fluctuations of both form and abundance of elemental components of proteins are occurring in natural habitats. One would thus expect that specific molecular mechanisms might have evolved in order to allow living organisms to respond to the elemental variations of the environment that they inevitably faced during their natural history. These adaptative processes undoubtedly left traces in the chemical composition of biopolymers and it could be anticipated that the advent of complete genomic sequences, combined with the use of straightforward statistical methods, would open the way to retrieve such imprints in macromolecules sequences.

For approaching this question, we recently investigated the atomic composition of proteins subsets from both the bacteria *Escherichia coli* and the eucaryote *Saccharomyces cerevisiae* [7]. Such a study allowed us to discern in both organisms, that the atomic composition of several protein families has been constrained in response to specific selective

pressures. A pronounced correlation between atomic compositions and metabolic roles were deciphered in enzymes involved in essential nutriment assimilation in these microbial organisms. For instance, carbon assimilatory enzymes were found to contain significantly less carbon atoms than the total proteins of *E. coli* and *S. cerevisiae* [7]. As well, sulfur assimilatory enzymes of both organisms were shown to comprise less sulfur atoms that the bulk proteins. Such impoverishments of sulphur and carbon assimilatory enzymes in their respective element were interpreted as an imprint of variations in the nutritional availability of these elements during the natural history of *E. coli* and *S. cerevisiae*. These results thus suggested that the evolution of proteins could have been more subjected to ecological constraints than previously thought. In a next step, we wanted to know how broad the fluctuations of elemental protein composition might be, and more specifically, whether the atomic contents of proteins could display substantial variations among different organisms.

Here we report the first results of the systematic analysis of the atomic contents of the proteins encoded by organisms whose genome has been entirely sequenced. Complete protein data sets from 46 organisms, representative of the bacteria, archaea and eucaryote kingdoms were compiled and analyzed using a robust statistical approach based on quantile distributions. The results show that the atomic composition of proteins, and more especially their carbon contents, widely differ among species and may differ more between species than within proteomes. Protein carbon content variations were further shown to apply to ortholog proteins : indeed their carbon contents vary among species as those of bulk proteins. A strongly significant correlation between the average carbon content of proteomes and the DNA base composition of the genomes was moreover uncovered. Randomisation methods and analysis of the genetic code further demonstrate that the extent genetic code relates DNA G+C usage with carbon content of proteins.

Materials and Methods

Organisms used for this study

Five eucaryotic, ten archaeal and thirty-one bacterial proteomes were downloaded from GenBank FTP site (ftp://ftp.ncbi.nih.gov/genbank/genomes/). All the proteins containing one or more undetermined amino acid ("X") were discarded, except for *Arabidopsis thaliana* proteins in which only the "X" amino acids were discarded. The coding G+C contents were retrieved from the Kazusa's Codon Usage Database (www.kazusa.or.jp/codon/) and the prokaryotic total G+C contents were found in the literature. The considered organisms are the following: *Arabidopsis thaliana* (25,531 proteins; G+C coding = 44.17%), *Caenorhabditis elegans* (17,074 proteins; G+C coding = 42.59%), *Drosophila melanogaster* (14,335 proteins; G+C coding = 53,99%), *Homo sapiens* (24,493 proteins; G+C coding = 52.51%), *Saccharomyces cerevisiae* (6,330 proteins; G+C coding = 39.72%), *Aeropyrum pernix* (2,694 proteins; G+C coding = 57.49%; G+C total = 56.3%), *Archaeglobus fulgidus* (2,420 proteins; G+C coding = 49.37%; G+C total = 48.5%), *Halobacterium sp*. (2,058 proteins; G+C coding = 65.26%; G+C total = 65.9%), *Methanococcus jannaschii* (1,773 proteins; G+C coding = 31.94%; G+C total = 31.09%), *Methanobacterium thermoautotrophicum* (1,869 proteins; G+C total = 49,50%), *Pyrococcus abyssi* (1,765 proteins; G+C coding = 45.16%; G+C total = 44.7%), *Pyrococcus horikoshii* (2,038 proteins; G+C coding = 42.32%; G+C total = 42.0%), *Sulfolobus solfataricus* (2,977 proteins; G+C coding = 36.30%), *Thermoplasma volcanium* (1,499 proteins), *Thermoplasma acidophilum* (1,478 proteins; G+C coding = 47.38%; G+C total = 46%), *Aquifex aeolicus* (1,521 proteins; G+C coding = 53.58%; G+C total = 43.4%), *Bacillus halodurans* (4,066 proteins; G+C coding = 44.33%; G+C total = 43.7%), *Bacillus subtilis* (4,100 proteins; G+C coding = 44.35%; G+C total = 43.5%), *Borrelia burgdorferi* (1,606 proteins; G+C coding = 29.33%; G+C total = 28.6%), *Campylobacter jejuni* (1,654 proteins; G+C coding = 30.98%; G+C total = 30.6%), *Caulobacter crescentus* (3,767 proteins; G+C coding = 67.42%; G+C total = 67.2%), *Chlamidia muridarum* (898 proteins; G+C coding = 40.64%; G+C total = 40.3%), *Chlamidophila pneumoniae* (1,092 proteins; G+C coding = 41.29%; G+C total = 40.6%), *Deinococcus radiodurans* (3,089 proteins;

G+C coding = 67.24%; G+C total = 66.6%), *Escherichia coli* (4,289 proteins; G+C coding = 51.11%; G+C total = 50.8%), *Haemophilus influenzae* (1,666 proteins; G+C coding=37.52%; G+C total=38.1%), *Helicobacter pylori* (1,565 proteins; G+C coding=40.45%; G+C total=39%), *Lactococcus lactis* (2,266 proteins; G+C coding=35.69%), *Mesorhizobium loti* (6,752 proteins; G+C coding=57.29%; G+C total=62.47%), *Mycobacterium leprae* (1,605 proteins; G+C coding=59.63%; G+C total=57.8%), *Mycobacterium tuberculosis* (4,137 proteins; G+C coding=65.79%; G+C total=65.6%), *Mycoplasma genitalium* (489 proteins; G+C coding=31.74%), *Mycoplasma pneumoniae* (689 proteins; G+C coding=41.06%; G+C total=40.0%), *Mycoplasma pulmonis* (782 proteins; G+C coding=26.38%), *Neisseira meningitidis* (2,081 proteins; G+C coding=51.49%; G+C total=51.5%), *Pasteurella meningitidis* (2,014 proteins; G+C coding=37.81%; G+C total=40.4%), *Pseudomonas aeruginosa* (5,565 proteins; G+C coding=66.63%; G+C total=66.6%), *Rickettsia prowazekii* (834 proteins; G+C coding=30.60%; G+C total=29.1%), *Staphylococcus aureus* (2,594 proteins; G+C coding=32.90%), *Streptococcus pyogenes* (1,696 proteins; G+C coding=40.99%; G+C total=38.5%), *Synechocystis sp.* (3,168 proteins; G+C coding=49.52%; G+C total=49.5%), *Thermotoga maritima* (1,849 proteins; G+C coding = 46.45%; G+C total = 46.2%), *Treponema pallidum* (1,003 proteins; G+C coding = 52.52%; G+C total = 52.8%), *Ureaplasma urealyticum* (613 proteins; G+C coding = 26.55%; G+C total = 25.5%), *Vibrio cholerae* (3,822 proteins; G+C coding = 47.62%; G+C total = 47.48%), *Xylella fastidiosa* (2,766 proteins; G+C coding = 53.78%; G+C total = 52.7%).

Orthologous proteins

Sequences of proteins comprised within 27 different clusters of orthologous proteins (COGs) spanning 42 species (bacteria or archaea) were retrieved (http://www.ncbi.nlm.nih.gov/COG/). The 27 used COGs were COG0006, COG0012, COG0013, COG0018, COG0024, COG0030, COG0050, COG0060 COG0080, COG0081, COG0090, COG0091, COG0093, COG0112, COG0130, COG0125, COG0197, COG0200, COG0201, COG0244, COG0256, COG0442, COG0459, COG0495, COG0541, COG0552, COG0575. Proteins belonging to these 27 COGs were clustered according to species or

according to COGs, the corresponding protein carbon usages were calculated and the resulting quantile distributions were displayed as for whole proteomes.

Random Proteomes

Random protein samples were first produced using the natural codon catalog from random DNAs displaying different G+C contents. For each G+C content value, five different protein samples were generated and their resulting carbon quantile distributions were calculated. Each protein sample is composed of 2000 random proteins the lengths of which are distributed among nine values, ranging from 100 to 600 residues, to mimic the distributions of real protein lengths. Codons were sorted out with the unique constraints of having $P_A = P_T$ and $P_C=P_G$, where Px is the probability of sorting the base X out of the set (A, C, T, G), and P_G is chosen to raise the wanted final G+C content. Stop codons were systematically discarded.

To analyze random protein samples produced with randomly generated genetic codes, 500 plausible alternative codes were first generated following the criteria of Haig and Hurst [8] which are: *i)* to assign one of the twenty amino acids to each cluster of codons coding for one amino acid in real life and *ii)* to keep the stop codons. From each randomly generated code, six randomly generated genomes coding for 2000 proteins of 280 residues were computed as described above and raised a final G+C content of respectively 28.7, 37.5, 44.4, 50.9, 59.6 and 65.6 %. The correlation factors r between those six G+C content values and the resulting carbon mean values were calculated for each code. The quantile distribution of the slope of the corresponding regression lines were plotted for all the codes leading to a squared correlation factor > 0.9.

Computerized DNA mutations

Randomization essays were performed starting from the DNA sequence of an *E.coli* gene coding for a protein of length L. Repeatedly, one position of the DNA sequence was randomly chosen and the corresponding base was proposed to be mutated into one randomly

chosen different base. Every position in the DNA sequence had the same probability to be chosen. Similarly, replacement with each of the three remaining base was equiprobable. Only mutations that would strictly decrease the carbon content of the encoded protein were kept. The number of effective mutations, the GC value of the resulting DNA sequence and the carbon content of the corresponding protein were checked every L fold proposals of mutation.

Results

Large variations of protein atomic contents in firmicute species.

To analyse the atomic content of proteins encoded by entirely sequenced genomes, we did not used average values but we applied more robust statistical methods, the quantile distribution method and stochastic ordering concepts [9, 10]. For each organism, the quantile distributions of carbon, nitrogen and sulfur atom frequencies within the residue side chains of the encoded proteins were calculated. The quantile Q^S_A (x) is the fraction of proteins from the specie "S" in which the averaged number of the atom "A" per residue side chain is at most "x". The medians (the Q_A^{-1} (0.50) point) and the 80% quantile range (corresponding to the 0.10- 0.90 quantile levels) are major statistical measurement of atomic contents [10].

As a first assessment of how atomic composition of proteins fluctuates on a proteome-wide level, we chose the phylogeneticaly related group of firmicutes, a bacteria family that comprises the well known model organism *Bacillus subtilis* and provides a large set of entirely sequenced genomes [11]. The results of the analysis made with 11 different firmicute bacteria are depicted in figure 1. As previously noted for *E. coli* and *S. cerevisiae* [7], the quantile distributions indicate that, for each organism, the carbon, nitrogen and sulfur contents of proteins follow bell-shaped distributions which are nearly Gaussian. Importantly, the results show that, depending on the atom considered, the atomic quantile distributions are differently scattered among species. Indeed, while quantile distributions for carbon, and to a lesser extent for sulfur, display substantial variations between species, the nitrogen quantile distributions are largely similar in all analyzed bacterial species. The statistical significance of the differences observed between carbon (or sulfur) distributions can be assessed by using normal z-tests. Two-sided P value ($<10^{-9}$) confirms that, for instance, the differences observed between the *Bacillus halodurans* and *Staphylococcus aureus* carbon distributions are significant and likely not to occur by chance. A simple hypothesis accounting for the results

reported in figure 1 would have been that the scattering of the quantile distributions seen among species reflects the variations of the number of the atom considered within the twenty canonical amino acids. The clustering of the nitrogen, but not of the sulfur quantile distribution, rules out this hypothesis. Indeed, sulfur protein contents broadly vary between firmicute species while amino acids contain at most one sulfur atom. In contrast, nitrogen protein contents are clustered although either 0, 1, 2 or 3 nitrogen atoms are found in the side chains of canonical amino acids.

Altogether, the results displayed in figure 1 suggest that, at the atomic level, the composition of the proteins encoded by each genome broadly varies from one organism to another and overemphasize at a proteome-wide level, the plasticity of protein atomic contents established for cyanobacterial light-harvesting proteins [12] and microbial assimilatory enzymes [7, 13].

Protein carbon contents differ more between species than within each proteome.

Since the larger variations in protein atomic content were observed for carbon, which is the main architectural component of proteins, we focused our analysis on this atom (sulfur variation analyses will be reported elsewhere). Figure 2A shows that the variations in protein carbon content observed for the firmicutes also hold for other bacteria species as well as for the archaea and eucaryote kingdoms. Remarkably, the carbon distributions for all kingdoms appear to be strictly ordered, with no (bacteria and archaea) or only few (eucaryotes) crosses between the plots, even at the outliers of the distributions. This indicates that while the mean values of protein carbon contents largely differ between species, the spread of each distribution is equivalent for all the species analyzed. As the overall dispersal of the carbon quantile distributions is larger than the spread of each quantile distribution (figure2A), the atomic compositions of proteins appears to be more different between species than within each organism. The stochastic ordering of the carbon quantile distributions observed for both archaea and bacteria species further demonstrates that at each level of protein carbon content,

the quantile points are similarly ranked between species. Thus, whatever the origin of the observed variations in protein carbon content was, their cause seems to have change uniformly the totality of the proteins encoded by each organism.

To further assess this point, we then analyzed the carbon contents of several ortholog proteins shared by archaea and bacteria species. Sequences of proteins comprised within different clusters of orthologous proteins (COGs, [14]) were retrieved from the COG database and the carbon contents of the retrieved proteins were calculated and analyzed. First, the quantile distributions of carbon usage of proteins belonging to 27 different COGs were calculated for each species. As depicted in figure 3A, the quantile distributions restricted to ortholog proteins are dispersed among species as the carbon quantile distributions calculated for all the proteins of each species. In addition, this analysis demonstrates that the quantile distributions of ortholog and total proteins are similarly ordered among species. Next, the carbon content of ortholog proteins was calculated as gathered by COGs. Carbon quantile distributions were calculated for each of the 27 COGs and compared to the quantile distribution of the averaged carbon content of whole proteomes. As depicted in figure 3B, the slopes of the quantile distributions of carbon contents of ortholog proteins gathered by COGs is similar to the slope of the quantile distribution of the median carbon values ($Q_C^{-1}(0.5)$) of total proteins. This shows that, within a COG, the carbon content of ortholog proteins vary among species as the averaged carbon content of whole proteomes. Thus, both results establish that carbon content of ortholog proteins vary among species as the bulk proteins.

These results are further striking as the uncovered biases in protein carbon contents are quantitatively significant. Calculations indeed reveal that the amount of carbon atoms needed for protein synthesis may differ from more than 12 % between species. This means that, depending on the organism, from 550 to 660 carbon atoms are, on the average, utilised to build the residue side chains of a protein of 200 residues. We noticed that, according to the compilation of Neidhardt [15], such a difference may account for more than 6 % of the totality of the carbon atoms that are fixed to construct a canonical bacterial cell such as *E. coli.*

Proteome carbon content is related to genome base composition

Taken together, all the above reported results strongly argue that the protein carbon content biases result from constraints that were superimposed to those acting on activity, folding and stability of proteins and which are generally view as the primary determinants of protein evolution [16, 17]. We thus searched for a plausible origin of these atomic composition biases. We first tried to establish whether variations of protein carbon contents result from the impoverishment or enrichment of particular amino acids. Protein amino acid composition of organisms displaying various mean protein carbon contents were retrieved and compared. Such a comparison, made with seven prokaryotes, showed that organisms with high protein carbon contents contain, on the average, more asparagine, lysine and isoleucine residues (figure 4). Conversely, organisms with low protein carbon content encode proteins enriched in glycine and alanine residues. The differences are however weak and the overall trend is rather that protein atomic variations result from variations in the usage of a large number of amino acids.

Various traits of the analysed species were next inspected but this did not suggest a direct explanation for the origin of the biases. However, we noticed that the ordering of the carbon distributions was somehow related to phylogeny and taxonomy. This pointed us at the guanosine plus cytosine composition of genomic DNAs which displays both large inter- and low intra-specific variations as the above reported protein carbon biases. Indeed, DNA G+C content is known to vary from approximately 25 % to 75 % between species, with closely related species generally having similar DNA G+C contents [18]. We thus examined whether the average protein carbon content might be correlated to DNA G+C content. As shown in figure 5A, a strong correlation between the genome nucleotide composition and the mean value of protein carbon contents was uncovered. The results are statistically highly significant, the median of the carbon quantile distributions being correlated with both the G+C content of the coding DNA ($P < 10^{-16}$, 44 species) and the G+C content of the total DNA

($P < 10^{-15}$, 40 species). In addition, regressions were calculated for each of the three kingdoms, bacteria, archaea and eucaryotes, confirming that this relation applies for every lineages (figure 5B). In contrast, there is no correlation between protein sulfur content and genome composition although protein sulfur distributions widely vary among species (figure 5A). To further assess that the correlation observed between the median quantile point of the carbon distributions and the G+C content indeed apply to the whole proteome and was not a result of extreme carbon usage in particular proteins or protein families, we used the 80% quantile ranges. Regressions between both the $Q_C^{-1}(0.10)$ and the $Q_C^{-1}(0.90)$, and the G+C DNA content were calculated and, as for the mean carbon content, highly significant correlations were found at both values (figure 5C). Finally, we compared the carbon content of ortholog proteins belonging to several COGs and their relationship to the G+C content of the genomes that encode them. As shown in Table 1, the relationships between low protein carbon content and high DNA G+C content hold as well for ortholog proteins. All these results therefore provide strong evidences that the carbon content of proteins is strongly related to the base composition of their encoding genome.

Proteome carbon content and genome base composition relationship is imprinted in the extent genetic code.

That proteome atomic composition is so closely entwined with genome base composition also suggests that this relation is imprinted in the structure of the canonical genetic code. To test the hypothesis that the genetic code indeed relates low protein carbon content to high DNA G+C content, we first followed a randomization approach. Using the natural codon sets, random protein samples were produced from random DNAs with different G+C contents (see Materials and Methods) and the resulting protein carbon quantile distributions were calculated. This shows that all the random protein samples produced from random DNAs having the same G+C content display the same carbon quantile distributions (figure 6A). Moreover, the random protein carbon distributions are stochastically ordered

according to the G+C content of the randomly generated DNAs. As shown in figure 6A, we found a strong correlation between nucleotide composition of random genomes and the median quantile value $Q_C^{-1}(0,5)$ of carbon contents of the corresponding proteins. On the contrary, protein sulfur content distributions of random proteomes do not correlate with the base composition of random DNAs (data not shown) as observed for real organisms.

To further analyse how the extent genetic code relates low proteome carbon content to high G+C genome content, we next examined, within the current codon catalog, all the single mutations that replace one amino acid by another and we determined how single mutations that lower or increase amino acid carbon content, in turn modify the G+C content of corresponding codons. Calculations (figure 6B) demonstrate that among the single mutations that lower amino acid carbon content, 49.05 % do increase G+C codon content, 35.22 % do not change the G+C codon content and only 15.72 % decrease the G+C codon content (as expected, the exact opposite trend was measured for mutations increasing amino acid carbon content with 49.05 % lowering and 15.72 % increasing the G+C codon content, respectively). Randomization assays show how the accumulation of mutations lowering the carbon content of a protein in turn increases the G+C content of its encoding gene (figure 6C).

Finally, we wondered whether the relationship that we evidenced above is specific to the extent genetic code. We therefore analyzed random protein samples produced with randomly generated codes. 500 plausible alternative codes were generated according to the criteria of Haig and Hurst [8]. From each randomly generated code, six randomly generated genomes coding for 2000 proteins of 280 residues and raising 6 different G+C contents were computed. The correlation factors r between those six G+C content values and the resulting carbon mean values were calculated for each code. The quantile distribution of the slopes of the corresponding regression lines were plotted for all the codes leading to a squared correlation factor > 0.9. As shown in figure 6D, the random reassignments of the 20 amino acids to the codon sets observed in the canonical genetic code show that less than 6 % of randomly generated codes allow protein carbon content to be negatively correlated with DNA G+C content and to vary with an amplitude equal or larger than what is observed with the

canonical genetic code. Therefore, the ability of relating low protein carbon content to high G+C DNA content appears to be a property shared by only few genetic codes that could be generated, following the criteria of Haig and Hurst [8], from the codon sets found in the canonical code.

Discussion

The here presented results reveal first the remarkable plasticity of the elemental composition of proteins, a striking but previously overlooked feature of this class of biopolymers. Indeed, it appears that not only the atomic composition of proteins may vary from one species to another, but the differences can be very large and widespread. The results show that this holds for two elemental components of proteins, sulphur and carbon, the larger variations being measured for the latter atom. Importantly, our analyzes revealed that differences in protein carbon, and to a lesser extent sulphur, contents are more pronounced between species than within species. Quantile analyzes further reveal that the differences in protein carbon content observed among organisms are not the results of particular biases in some specific protein subclasses but apply to the entire content of each proteome. Specific calculations made with ortholog proteins confirm this assertion and show that the carbon content of ortholog proteins varies as the carbon content of bulk proteins. This clearly indicates that the evolution of protein atomic structures have been influenced by constraint other that the ones acting on activity, specificity, folding and stability. Our results further generalize at a proteome-wide level the variation of protein atomic contents already established for several subsets of enzymes such as cyanobacterial light-harvesting proteins [12] and microbial assimilatory enzymes [7]. The here uncovered feature of protein evolution at a proteome-wide level certainly escaped previous attention as most of the studies devoted to protein evolution focused on amino acid composition and as carbon content differences does not result in biases of one or few specific amino acids but into variations in the usage of most amino acids. It is worth to note that the biases in protein carbon contents observed between species could be quantitatively significant as they could account for more than 6 % of the carbon mass which is required to construct a microbial cell. As large fluctuations of carbon abundance are known to occur in natural habitats [6], this latter result strengthens the possibility of nutritional constraints as shaping protein structures .

Looking for a plausible origin of such elemental biases of protein composition, we uncovered a strong correlation between protein carbon content and the base composition of the genomes. A highly significant correlation was indeed demonstrated, low protein carbon content being correlated with high DNA G+C content. This correlation was shown to exist in each of the three kingdoms, bacteria, archaea and eukaryotes and was demonstrated to be not a result of extreme carbon usage in particular proteins or protein families but to apply to entire proteomes. The analysis of several ortholog proteins belonging to different COGs shows that the relationships between low protein carbon content and high DNA G+C content hold as well for ortholog proteins. The correlation between DNA G+C and protein carbon contents strongly sustains the possibility that the composition of one class of biopolymers might have been determined by the other through the canonical genetic code. At the same time, this strong relation poses the conundrum of the flow of causality : DNA composition dictating protein atomic composition or the converse ?

The first mode (DNA composition dictating protein atomic composition), is in the line with the neutral theory of evolution [19], and more specifically with the original view of Suekoa, according to which genome composition biases are due to differences between the forward and backward mutation rates of the GC and AT pairs [3, 20]. Although the molecular basis of biased AT/GC pressure is unknown, it was suggested that it has been acted uniformly on all of the DNA [18]. Biased mutation pressure was proposed to result in turn into amino acid composition biases in proteins [4, 20, 21]. In this model, any modification in DNA base usage will gradually lead, along generations, to either frugal or wasteful carbon fixation. It has been noticed early that the smallest amino acids were indeed encoded by codons comprising the G and C bases only [22], yet the close relationship that exists between protein carbon and DNA G+C contents might have been overlooked.

The second mode (protein composition dictating DNA composition) fits with a more selectionist view of evolution and with the previous proposals that metabolic flows and geochemical budgets might be constraints that were imprinted on protein evolution [7, 23]. In a simple model, the adaptation to nutritional resources scarce in carbon could result in the

fixation of mutations trimming the protein carbon content, which in turn lead to a progressive enrichment of genomes in GC bases. Examination of the structure of the genetic code provides further arguments for such a model. Indeed, calculations and randomisation assays show how, given the structure of the canonical codon catalog, the successive accumulation of mutations lowering the carbon content of proteins will in turn lead to the enrichment in GC bases of the genomes. The fact that composition constraints may have played a role during the evolution of proteins have been previously delineated in the case of highly expressed proteins in cyanobacteria : the sulphur-oligotrophic Calothrix Sp PCC7601 encodes a sulphur-depleted version of its most abundant protein (phycocyanin) which it specifically expresses under environmental condition of sulphur limitation [12]. Likewise, Fauchon *et al.* recently reported how yeast cells, when exposed to a toxic metal (cadmium), save sulphur by reducing the expression of sulphur rich-proteins. In particular these authors demonstrated how, upon cadmium exposition, several abundant glycolityc enzymes are replaced by sulphur depleted isoenzymes [13]. Taken together, our results thus suggest that DNA G+C content biases might be a direct consequence of the optimisation of protein atomic contents in response to carbon availability in natural habitats. In this regards, it is interesting to notice that G+C contents exhibit a much smaller variation in metazoan species and, especially, that vertebrate genomes show quite a uniform averaged G+C content, ranging from 40% to 45 % [20, 24]. This could be accounted for by the fact that vertebrates are expected to be less subjected to a specific elemental nutritional constraint than micro-organisms. As well, free living bacteria tends to have genomes with significantly higher G+C contents, and thus proteins constructed with less carbon atoms contents, than pathogen or symbiont bacteria which rely on nutritional resources provided by their hosts and thus are expected to be less subjected to environmental carbon limitations [25]. Experimental confirmation of the hypothesis that genome base composition is driven by nutritional constraints shaping protein atomic contents would thus provide a new way to explore the origin of genome composition biases, a question that lies at the heart of current molecular evolutionary debates.

Acknowledgements

We are especially grateful to Bruno Sargueil and Rupert Mutzel for fruitful discussions and suggestions. P. B.-C. is supported by a thesis fellowship from the Ministère de la Défense.

References

1. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299-304.
2. Koonin, E. V., Aravind, L. & Kondrashov, A. S. (2000). The impact of comparative genomics on our understanding of evolution. *Cell* 101, 573-576.
3. Suekoa, N. (1961). Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb. Symp. Quant. Biol.* 26, 35-43.
4. Singer, G. A. C. & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* 17, 1581-1588.
5. Thorne, J. L. (2000). Models of protein sequence evolution and their applications. *Curr Opin Genet Dev* 10, 602-5.
6. Brock, D. & Madigan, M. T. (1991). *Biology of Microorganisms*, Prentice-Hall International, London.
7. Baudouin-Cornu, P., Surdin-Kerjan, Y., Marliere, P. & Thomas, D. (2001). Molecular evolution of protein atomic composition. *Science* 293, 297-300.
8. Haig, D. & Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. *J Mol Evol* 33, 412-417.
9. Karlin, S., Blaisdell, B. E. & Bucher, P. (1992). Quantile distribution of amino usage in protein classes. *Protein. Eng.* 5, 729-738.
10. Karlin, S. & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science* 257, 39-49.
11. Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* 51, 221-271.
12. Mazel, D. & Marlière, P. (1989). Adaptative eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature* 341, 245-248.
13. Fauchon, L., Lagniel, G., Aude, J. C., Lombardia, L., Soularue, P., Petat, C., Marguerie, G., Sentenac, A., Werner, M. & Labarre, J. (2002). Sulfur-sparing in the yeast proteome in response to sulfur demand. *Molec. Cell* 9, 713-723.
14. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631-637.
15. Kredich, N. M. (1996). Biosynthesis of cysteine. In *Escherichia coli and Salmonella typhymurium.* (Niedhardt, F. C., ed.), pp. 419-. American Society for Microbiology, Washington.

16. Nei, M. (1987). Molecular evolutionnary genetics. *Columbia University Press, New York*.
17. Tourasse, N. J. & Li, W. H. (2000). Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* 17, 656-664.
18. Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* 84, 166-169.
19. Kimura, M. (1983). The neutral theory of molecular evolution. *Cambridge University Press, Cambridge, england*.
20. Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* 48, 582-592.
21. Lobry, J. R. (1997). Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205, 309-316.
22. Woese, C. R. (1965). Order in the genetic code. *Proc Natl Acad Sci U S A* 54, 71-75.
23. Akashi, H. & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci U S A* 99, 3695-3700.
24. Bernardi, G. & Bernardi, G. (1985). Codon usage and genome composition. *J Mol Evol* 22, 363-365.
25. Rocha, E. P. & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet* 18, 291-294.

Legends to figures

Figure 1. Quantile representations of protein atomic content in the firmicute family of bacteria. For each proteome, the averaged number of either carbon, nitrogen or sulfur atoms found in residue side chains for each protein was calculated, and the totality of all these frequencies was described by a histogram. The quantile distributions are the cumulative representations of these histograms. For each protein sample, the quantiles were calculated so as to display the distribution by a 50 dot graph. The quantiles for *E. coli* were included as reference in each graph.

Figure 2. Protein carbon usage in bacteria, archaea and eucaryotes. The quantile distributions were calculated and are displayed as in figure 1. For bacteria, the carbon distribution plot does not include the firmicute distributions which are displayed in figure 1. As in Fig 1, the quantiles for *E. coli* were included in each graph.

Figure 3. A) Carbon quantile distributions of ortholog proteins shared by bacteria and archea species. The carbon usage of proteins belonging to 27 different clusters of orthologous proteins (COGs, [14]) was calculated for each indicated species and the corresponding quantile distribution were displayed. **B)** Carbon contents of ortholog proteins belonging to 11 different COGs were gathered by COG. The resulting COG carbon quantile distributions were displayed together with the quantile distribution of the median carbon contents ($Q_C^{-1}(0.5)$) of archaea and bacteria total proteins (plain line).

Figure 4. Amino acid composition of proteomes with different carbon contents. The averaged amino acid content of proteins encoded by seven different prokaryotic species exhibiting

different protein carbon contents was calculated from the Kazusa's Codon Usage Database (www.kazusa.or.jp/codon/) and depicted as rings. Species (*U. urealyticum*, *T. maritima*, *T. pallidum, M. tuberculosis, M. pneumoniae, D. radiodurans and B. subtilis*) were ordered outside to inside according to their averaged protein carbon contents. Amino acid were depicted in the single letter nomenclature together with the number of carbon atom found in their side chain.

Figure 5. Correlation between G+C content and protein carbon content. **A)** Mean protein carbon content decreases with increasing G+C content of genome DNA while no correlation is observed between either mean sulfur or mean nitrogen content of proteins and the G+C content of genome DNA. **B)** The correlation between G+C content and the mean protein carbon content applies in the three lineages, bacteria, archaea and eucaryotes. **C)** Regressions calculated for the $Q_C^{-1}(0.10)$ and $Q_C^{-1}(0.90)$ show that the correlation between genome G+C contents and protein carbon contents apply for at least 80 % of the total proteins encoded by each species.

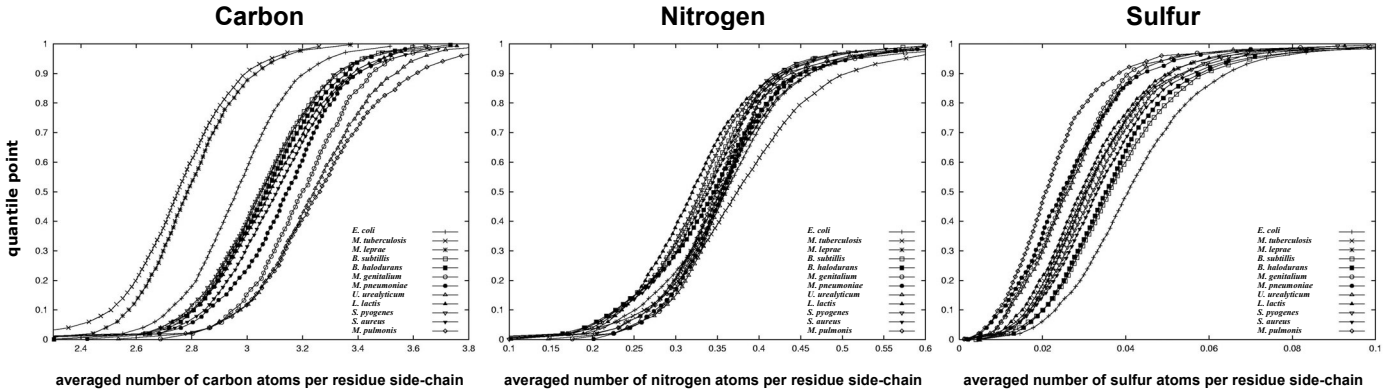
Figure 6. A) DNA G+C content correlates with protein carbon content in random samples generated using the canonical genetic code. For each G+C value, five different random protein samples were produced, each comprising 2000 proteins. The carbon content of the protein data sets was next analyzed by the quantile method and the corresponding distributions were plotted. For each G+C content value, the carbon quantile distributions of five different randomly generated protein samples are plotted (left panel). Correlation between G+C content and protein carbon content in the case of real proteomes (diamonds) and of randomly generated protein samples (circles, right panel) **B)** Distribution, in the 2D space of the codon G+C content changings and the corresponding amino acids carbon content

changings, of all the single point mutations transforming one amino acid into one another. Up and down arrows indicate increase and decrease, respectively, of amino acid carbon content and codon G+C content. **C)** Simulation of the evolution of protein carbon and gene G+C contents. The simulation was done using the *E. coli* ribosomal protein L1 (234 amino acids) as starting point and submitting its corresponding gene to computerized mutations. Five independent simulations, each represented by one curve, were performed. For each simulation, up to 1638 mutations were proposed and only those that strictly decreased the protein carbon content were kept. Carbon and G+C contents were checked every 234 proposed mutations (one mutagenesis round). **D)** Analysis of G+C content and protein carbon content correlation with 500 randomly generated genetic codes. The random genetic codes were generated using the method of Haig and Hurst [8]. For each randomly generated genetic code, six different genomes were randomly generated in order to have different G+C value contents. The correlation factors r between those six G+C content values and the resulting carbon mean values were next calculated. The graph displays the quantile distribution of the regression line slopes for all the codes leading to a squared correlation factor > 0.9. The position of the canonical genetic code is indicated. Only the codes which are situated at the left of the canonical genetic code within the quantile distribution allow protein carbon content to vary in response to G+C content with an amplitude equal or larger than what is observed with the canonical genetic code.

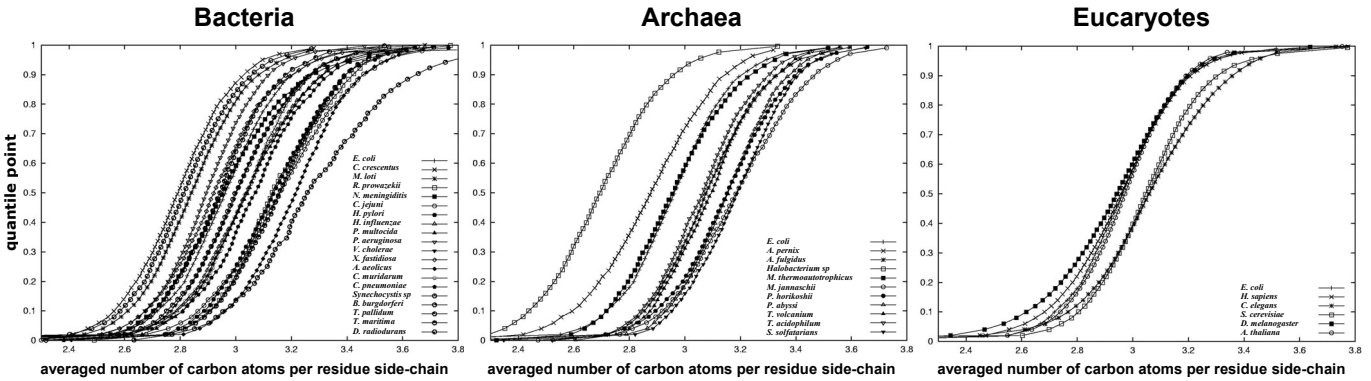
Table 1

COG		0012	0018	0030	0051	0081	0149	0575
Protein		Predicted GTPase	Arginyl-tRNA synthetase	Dimethyl-adenosine transferase	Ribosomal protein S10	Ribosomal protein L1	Triose-phosphate isomerase	CDP-diglyceride synthetase
E. coli gene		*ychF*	*argS*	*ksgA*	rpsJ	*rplA*	*tpiA*	*cdsA*
C. crescentus	residues	366	600	258	102	229	253	275
(G+C = 67,42 %)	carbon (average)	2,78	2,89	2,66	3.04	2,56	2,46	2,96
	carbon (total)	1752	2938	1204	514	1046	1130	1366
M. tuberculosis	residues	357	550	317	101	235	261	306
(G+C = 65,79 %)	carbon (average)	2,74	2,79	2,79	2.99	2,63	2,64	2,81
	carbon (total)	1695	2634	1520	504	1089	1213	1474
M. leprae	residues	356	550	306	101	235	261	312
(G+C = 59,63 %)	carbon (average)	2,76	2,78	2,87	3.00	2,66	2,68	2,83
	Carbon (total)	1696	2632	1493	505	1096	1223	1508
E. coli	residues	363	577	273	103	234	255	249
(G+C = 51,11 %)	carbon (average)	2,84	2,98	2,97	2.99	2,63	2,65	3,25
	carbon (total)	1760	2874	1357	514	1081	1187	1307
M. pneumoniae	residues	362	537	263	108	226	244	395
(G+C = 41,06 %)	carbon (average)	3,06	3,28	3,20	3.10	2,84	2,99	3,36
	carbon (total)	1834	2834	1368	551	1095	1218	2121
M.genitalium	residues	367	537	259	106	226	244	374
(G+C = 31,74 %)	carbon (average)	3,10	3,31	3,34	3.13	2,92	3,03	3,33
	carbon (total)	1871	2852	1385	544	1112	1228	1995

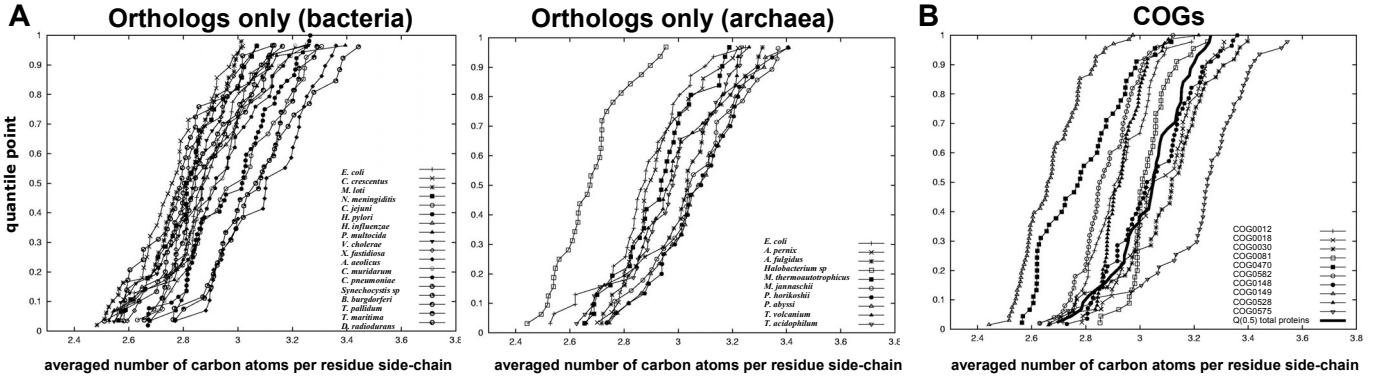
Legend : Variations, among six organisms displaying different DNA G+C contents, of the amount of carbon required for the building of several ortholog proteins. Ortholog proteins are defined according to the Clusters of Orthologous Proteins established by Tatusov *et al.* [14] . "average carbon" is the averaged number of carbon atoms found per residue side-chain, "total carbon" is the total number of carbon atoms used to build each protein.



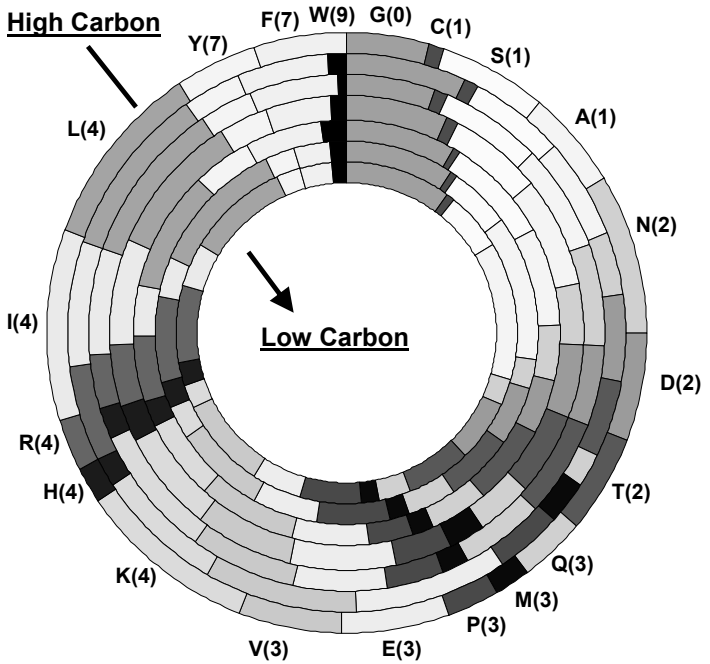
Baudouin-Cornu *et al.* **Figure 1**



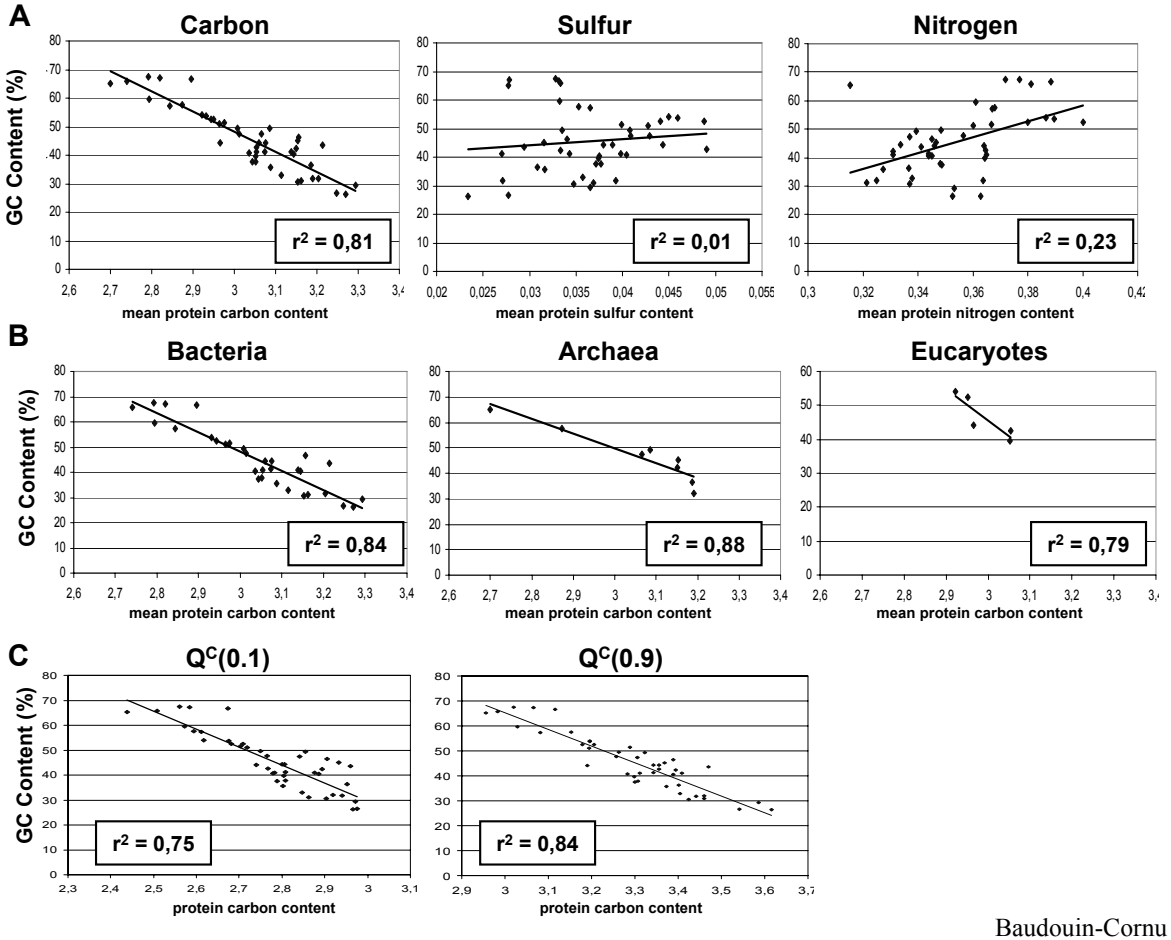
Baudouin-Cornu *et al.* **Figure 2**



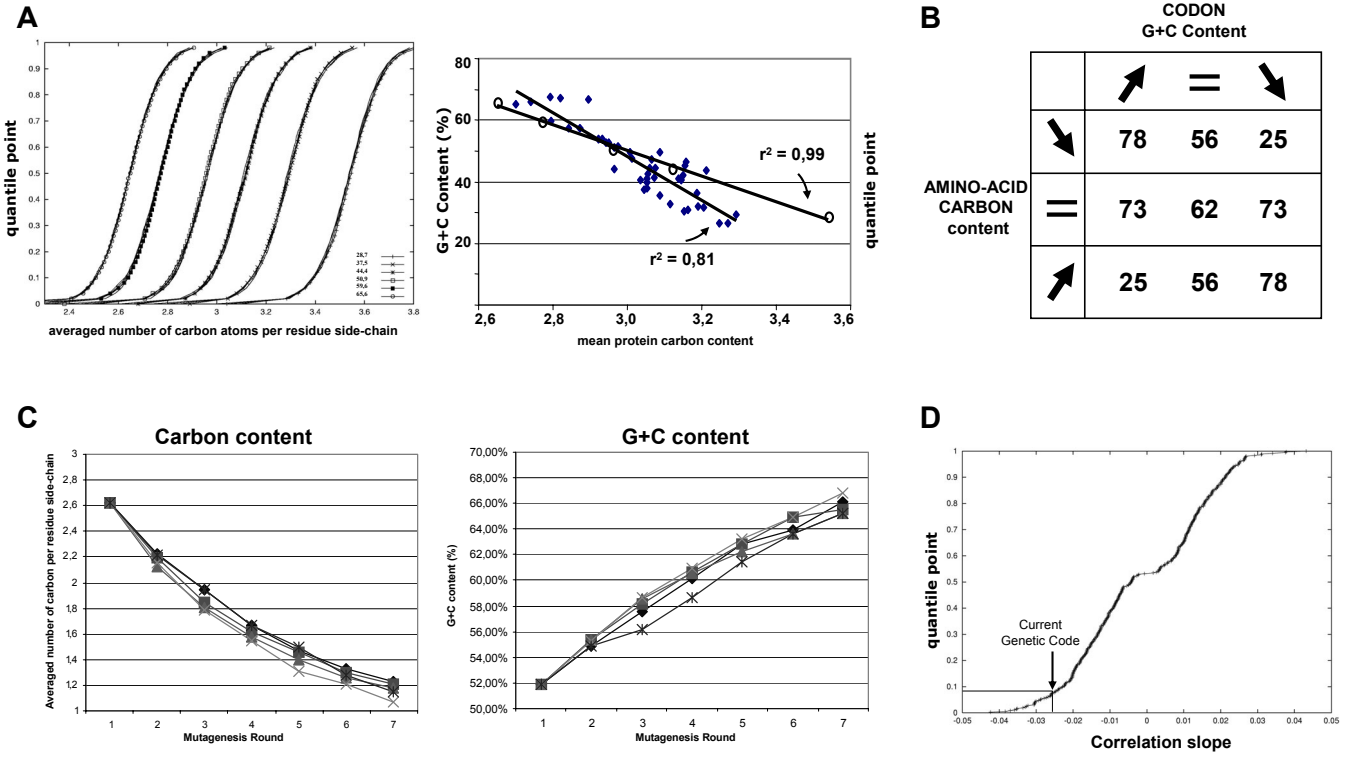
Baudouin-Cornu *et al.* **Figure 3**



Baudouin-Cornu *et al.* **Figure 4**



Baudouin-Cornu *et al.* **Figure 5**



Baudouin-Cornu *et al.* **Figure 6**